(Check) 24 - A Study and Model of Automatic Mapper For

by Arta Moro Sundjaja

FILE	24_-_A_STUDY_AND_MODEL_OF_AUTOMATIC_MAPPER_FOR.PDF (258.54K)		
TIME SUBMITTED	06-APR-2017 03:56PM	WORD COUNT	3633
SUBMISSION ID	649644345	CHARACTER COUNT	19959

A Study and Model of Automatic Mapper For ICD-9 and ICD-10 For Healthcare Provider in Indonesia

Yohan Muliono
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
Email: ymuliono@binus.edu

Ida S. Rejeki Siahaan
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
Email: isiahaan@binus.edu

Suharjito
Master in Computer Science
Bina Nusantara University
Jakarta, Indonesia
Email: suharjito@binus.edu

***Abstract*—Diagnosis Related Group (DRG) is a system of payment for classifying same clinical path of diagnosis and procedure into groups with similar payment. In DRG, there are some codes that being used called International Classification and Disease 9 and 10 (ICD-9 and ICD-10). DRG had already been adopted by many countries since 1983 and in Indonesia it is known as BPJS Kesehatan. BPJS Kesehatan has an online system, however many doctors still write medical records manually. Furthermore, there is a medical record staff called *coder* whose job is to map the handwriting of doctor into ICD-9 and ICD-10. Manual encoding may introduced errors due to weariness or misinterpretation of doctor's handwriting. Such a slight mistake leads to many disadvantages. One of the disadvantages is miscalculation of price claimed. This research is conducted to find a model to replace a manual process into automatic process using Natural Language Processing (NLP). Processes involved in this work are *term extraction* to extract medical terms, *term verification* to verify validity of terms extracted and *term mapping* to map Medical Record Status to ICD-9 and ICD-10 Code.**

***Index Terms*—Diagnosis Related Group; term extraction; ICD-9 and ICD-10 encoding; BPJS Kesehatan**

I. Introduction

DRG has been known worldwide since 1980. In 1983, United State government applied DRG at its healthcare provider. In 1987, 15 Europe countries applied DRG as well, followed by eight countries of commonwealth in 1988 [1]. In 2008, Indonesia started using DRG and well known as INA-DRGs handled by 3M (a provider of Grouper software). In 2010, INA-CBG t[illegible] is used to replace INA-DRG as the contract with 3M has ended and was replaced by the use of another grouper software company (University of United Nation Grouper UNU Grouper) [2].

Currently the INA-CBG are encoded manually by a medical record s[illegible]alled coder. The medical record being encoded contains the International Classification of Diseases-9 (ICD-9) for procedure and International Classification of Diseases-10 for (ICD-10) for diagnosis.

There are four parts in INA-CBGs code separated by "-", the structures of INA-CBG based on [3] code is as follows:

- First part is a group in Casemix Main Groups (CMG).
- Second part is case type (will be detailed in 2.3.2).
- Third part is spesific case of INA-CBGs (will be detailed in 2.3.1).
- Fourth part is severity level as Roman numeral system (I is not severe to IV is for very severe).

In INA-CBG, the translation process from medical record to ICD-9 and ICD-10 is still done by manual. Currently, healthcare providers in Indonesia are not used to electronic record for medical record. Therefore, doctor still writes medical records manually. And there is a medical record staff called "coder" whose job is to map the handwriting of doctor into ICD-9 and ICD-10. The drawback of manual encoding can be as follow: For instance, coders need to read a number of medical records. Therefore, their eyes may feel tired or they could be misinterpret the doctor handwriting. Such a slight mistake leads to many disadvantages. One of the disadvantages is miscalculation of price claimed.

This research focuses on the semantic process based on Natural Language Processing, [illegible]cially in Term Extraction and Term Mapping Method. Natural Language Processing (NLP) is a computer science field of study to make a program that can understand human language / natural language and analyze them. NLP has a primary role, processing a database or retrieving information from a large text [4]. NLP researchers usually want their machine to understand their language and manipulate them into their desired tasks [5].

The input of this research will be the medical record status S.O.A.P. (Subjective, Objective, [illegible]ssment and Plan), and the output is suggestion code for ICD-9 and ICD-10. Subjective contains the purpose of patient when he/she visit the hospital. Objective contains observation of the doctor to the patient For Example: heartbeat and temperature. Assessment contains complete diagnosis of patient. And Plan contains the medicine and treatment given to the patient. Medical Record Status and Medical Record have significant differences. Medical Record Status is a complete status of a patient contains S.O.A.P. and the patient track when being hospitalized whereas, Medical Record is only a summary of diagnosis and the ICDs.

978-1-5090-5130-4/16/$31.00 ©2016 IEEE

II. Literature Reviews and Related Works

A. S-Grouper

The main idea of this research is based on SGrouper ([6]). In 2015, [6] conducted a research on Semantic Grouper in Medical Record in Italy. They conducted a semantic grouper to translate medical record to ICD-9. The input of SGrouper is medical record by physicians and the output is suggestion of ICD-9 categories. This research has several contributions. First, Physicians can use this to improve the productivity. Second, Back office can use this to prevent the errors made by physicians and for the public administrator or local government they can compare the DRG codes identified by physicians with the semantic grouper. Our research improves their work. First, the language is different. SGrouper had ben developed in Italian language, while this research will be in Bahasa Indonesia. Second, the code suggestions. SGrouper made suggestion only for ICD-9, while this research will give both ICD-10 and ICD-9.

B. Diagnosis-Related Group

The adoption of DRG made a change from provider retrospective payment (PRP) system to prospective payment system (PPS) that is useful for improving hospital services. Adoption of DRGs based PPS has become an international trend since 1980s [1]. Miranda and Cortez [7] define DRG as a patient classification system that group patients according to the consumption of resources of treatment given and their clinical characteristics. DRG was first developed and adopted by Health Care Finance Administration (HCFA) to pay American hospitals in 1983 under the Medicare Program.

DRG had been adopted in European countries [1]. For example DRG system in Finlandia was adopted to assess hospital case-mix. Some countries adopted DRG and modify the payment system before implementing it in their countries. Whereas others only use DRG as their payment system without any modification. However the intention is still similar, that is to improve transparency, efficiency and to improve quality of hospital services.

DRG has been widely adopted across the globe not only by developed countries but also by developing countries for example: Vietnam and Indonesia. Countries which had adopted DRG believed that DRG based PPS could help them improve the old payment system. The reimbursement fees for DRG are in lump sum and the fees are determined based on average actual cost of each DRG case. Indonesia have been using DRG system since 2007 as INA-DRGs, latter INA-CBGs to improve the efficiency in the hospital sector.

C. The International Classification of Diseases

According to WHO, The International Classification of Diseases (ICD) are designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics [8]. In Indonesia ICD-9 is used as classification of procedures and ICD-10 is used as classification of diseases. There are several revisions of ICD, currently Indonesia using ICD version 2008.

TABLE I
Disease Category in ICD-10 [9]

Chapter	Disease Category
I	Certain infectious and parasitic diseases
II	Neoplasms
III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	Endocrine, nutritional and metabolic diseases
V	Mental, Behavioral and Neurodevelopmental disorders
VI	Diseases of the nervous system
VII	Diseases of the eye and adnexa
VIII	Diseases of the ear and mastoid process
IX	Diseases of the circulatory system
X	Diseases of the respiratory system
XI	Diseases of the digestive system
XII	Diseases of the skin and subcutaneous tissue
XIII	Diseases of the musculoskeletal system and connective tissue
XIV	Diseases of the genitourinary system
XV	Pregnancy, childbirth and the puerperium
XVI	Certain conditions originating in the perinatal period
XVII	Congenital malformations, deformations and chromosomal abnormalities
XVIII	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S Injury, poisoning and certain other consequences of external causes
XX	External causes of morbidity
XXI	Factors influencing health status and contact with health services

1) International Classification of Diseases-9 (ICD-9): ICD-9 format has 3, 4, or 5 digits. Codes with three digits are included in ICD-9 as the heading of a category of codes that may be further subdivided by the use of fourth and/or fifth digits, which provide greater detail. A three-digit code is to be used only if can not be further divided. Where fourth-digit subcategories and/or fifth-digit sub-classifications are provided, then they must be assigned. A code is invalid if it has not been coded to the full number of digits required for that code. For example, Acute myocardial infarction, code 410, has fourth digits that describe the location of the infarction For example, 410.2, of inferolateral wall [8] .

2) International Classification of Diseases-10 (ICD-10): As stated in [9], ICD-10 is structured as three to four characters. The first character of the ICD code is a letter, and each letter is associated with a certain disease. The list of chapter will be listed in Table I. Next characters is specified with numbers defined by the detail of diseases occurred in first characters. More numbers giving more detail to the disease. For Example:

- A00-B99 is Certain infectious and parasitic diseases
- A15-A19 Tuberculosis
- A15 is Respiratory tuberculosis
- A17 is Tuberculosis of nervous system
- A18 is Tuberculosis of other organs
- A19 is Miliary tuberculosis

D. Term Extraction

In 2000, Frantzi, Ananiadou and Mima conducted a research to extract terms in medical corpus [10], their research involved a method called C-Value and NC-Value. C-Value calculates a function to extract multi-word (more than 1 word). NC-Value calculates a function of co-occurrence of term using C-Value. A corpus preparation is done before calculating C-Value including Part-Of-Speech Tagging (POS Tagging), linguistic filter for each N-gram word, and the stopword filter. This method is very powerful for extracting multi-word terms, but not for single words, because the formula being used to calculate the C-Value cannot be applied for word with length one. To overcome the problem Nakagawa and Mori [11] conducted a research to calculate the C-Value for single word modifying [10] and named it MC-Value which stands for Modified C-Value. C-Value has been used by many researchers to help them in many contexts in recent research like retrieving terms [12], text mining [13] and optimized C-Value with genetic algorithm [14].

$$MC - value(a) = length(a)(n(a) - \frac{t(a)}{c(a)}) \quad (1)$$

a : candidate word
$length(a)$: is the number of single-nouns which make up candidate word
$n(a)$: total frequency of occurrence of a on the corpus
$t(a)$: frequency of occurrence of a in longer candidate terms
$c(a)$: number of those candidate terms

Formula for C-Value is used if the word is not a single word :
if the word is not nested

$$C - value(a) = log_n |a| .f(a) \quad (2)$$

else

$$C - value(a) = log_n |a| (f(a) - \frac{1}{P(T_a)} \sum_{b \epsilon T_a} f(b)) \quad (3)$$

a : candidate string
f : is frequency of occurence in corpus
T_a : set of extracted candidate terms that contain a
$P(T_a)$: number of these candidate terms

1) TF-IDF: In 1983, a famous scheme of term extraction was proposed and named term frequency inverse document frequency (tf idf) [15]. The basic vocabulary of "words" or "terms" are chosen, for documents in the corpus and a count is formed on the number of occurrences of each word. tf.idf scheme is still being used an famous in this era as term classification technique to verify whether a term is important or just a normal word used in many or even every document. Tf idf is an algorithm where terms will be weighted depend on its occurences in documents. If a term is occur in many documents, a score for that term weight will be low. And if a term is occur frequently in single or few documents, that term weight will be high.

$$W_{i,j} = tf_{i,j}.log\frac{N}{N_i} \quad (4)$$

and

$$tf_{i,j} = \frac{f_{ij}}{max\{f_{1j}, f_{2j}, ..., f_{|v|j}\}} \quad (5)$$

$W_{i,j}$ weight assigned to term i document j
$tf_{i,j}$ number of occurrence of term i in document j
N number of documents in entire collection
n_i number of documents with term i

In 2009, Branden et al., [16] performed a research on Integrating case-based reasoning system with an electronic patient record. They used Genetic Algorithm for Excelicare to perform feature weighting. Unfortunately, the application only supported similarity assessments, and they were still developing the system management interface to allow configuration of parameters, experimentation and case maintenance. In United Kingdom a big portion of patients data has been arranged neatly in electronic patient record, therefore this way of thinking will not be working in some countries.

In 2012, [17] conducted a research on DRG Classification for acute myocardial infarction, the purpose of the research is to assess classification variables and use certain algorithm to group patients with Acute myocardial infarction disease into DRG, and make the quasi price for each of them. They are using data from 11 countries. Including : Austria, England, Estonia, Finland, France, Germany, Ireland, Netherlands, Poland, Spain and Sweden. The data required are being used for making a quasi price for each country for each DRG classified from the Acute myocardial infarction disease, which is converted from national measures of DRG weight (i.e. cost weight, average tariffs, scores-taking account of outlier deduction/add-ons or additional payment where possible) using national conversion rates.

In 2012, [18] conducted a research to designed an application for German and English data and constructed a tool for monolingual term candidate extraction. Using C-Value as pre-processing phase of term extraction. In 2013, Hsien-Tseng Wang and Abdulla Uz Tansel [19] conducted a research on solving a different paradigm for case-based reasoning using either implicit or explicit aspects of the knowledge. This research used ontology features to support distributed case-based reasoning systems in medical decision support. [19] applied UMLS (Unified Medical Language System) to help interpretation and understanding of medical meanings across application systems. Gene Ontology, and SNOMED-CT (Systematized nomenclature of medicine clinical terms) as a multiaxial coding system.

III. Methodology

A. Data Collection Method

There are several data collection methods used in this research:

- **Literature Studies**, this research is based on Natural Language Processing, especially Term Extraction and Term Mapping. Publications were collected to be studied and searched through appropriate algorithm to fit in this research, so that the purpose of this research will be fulfilled.
- **Interview**, this research conducts several interviews with experts (doctors and medical record staffs) for gathering knowledge to complete this research.
- **Presentation**, this research outline has been presented in together with two other collaborators to Jaminan Kesehatan Nasional (JKN). JKN is government program which is intended to give a health insurance to Indonesian citizen. From the presentation, some inputs were collected to help completing this research.
- **Manual Input: from Doctor Handwriting to Electronic Record**, this research needs electronic medical record as testing and learning data set, we had already requested medical record data from several hospitals. However, due to data privacy issue, we are only able to obtain data from a private-small-scale hospital in West Jakarta. This hospital is willing to give the medical record with a Non Disclosure Agreement letter, but their medical records are still handwritten. To fulfil the need of this research, the handwritten data is typed to electronic records to be processed with the algorithm proposed in this research. The number of medical records converted will be 650 records.

B. Proposed Model

In current system, some doctors write medical records manually and give them to medical record staffs to be translated manually into ICD-9 and ICD-10 code. This research proposed a model to change a manual process of translation by medical record staff such that a doctor directly entries medical record status S.O.A.P. into system or a doctor still gives the medical record status to medical record staff to input the medical record status S.O.A.P. and the output will be suggestion for ICD-9 and ICD-10. In this case, human error in translating ICD-9 and ICD-10 code will be reduced.

C. Evaluation Method

Both training and testing data will be evaluated which is 650 medical record status S.O.A.P. in total. This research mostly focused on improving the term extraction method, in case of that, evaluation method will be compared with [10]. Comparison will be using precision and accuracy method. Precision is computed as in equation 6 and accuracy is computed as in equation 7 which both are based on [20].

$$Precision = \frac{TP}{Actual\ Positive} \tag{6}$$

$$Accuracy = \frac{TP + TN}{Total\ Data} \tag{7}$$

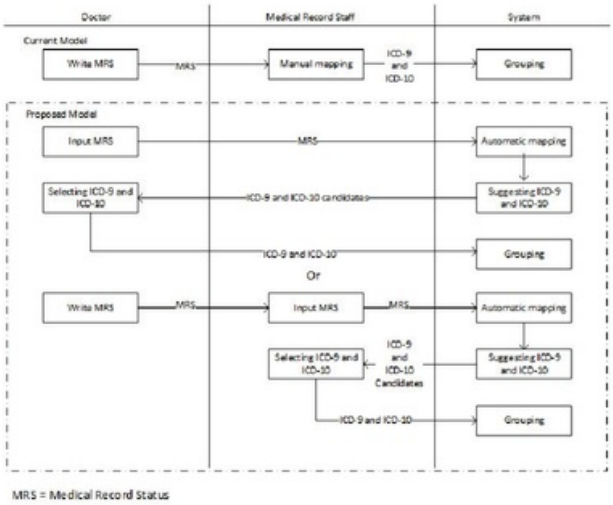

Fig. 1. Current Model and Proposed Model

$TruePositive(TP)$: Predicted True and Actual True
$TrueNegative(TN)$: Predicted False and Actual False
$FalsePositive(TN)$: Predicted True but Actual False
$FalseNegative(FN)$: Predicted False but Actual True

1) Program Evaluation Method: Result of this algorithm will be evaluated using [10], because technically this program is using statistical method and algorithm from [10]. The differences are our work is in Bahasa Indonesia and we justify the linguistic grammar rules used.

IV. Conclusion

Diagnosis Related Group (DRG) is a system of payment for classifying same clinical path of diagnosis and procedure into groups with similar payment. DRG had already been adopted in Indonesia in BPJS Kesehatan. BPJS Kesehatan has an online system and a medical record staff called *coder* whose job is to map the handwriting of doctor into ICD-9 and ICD-10. Manual encoding may introduced errors due to weariness or misinterpretation of doctor's handwriting. This research proposed a model to replace a manual process into automatic process using NLP. Processes involved in this work are *term extraction* to extract medical terms, *term verification* to verify validity of terms extracted and *term mapping* to map Medical Record Status to ICD-9 and ICD-10 Code.

References

[1] R. Busse, A. Geissler, A. Aaviksoo, F. Cots, U. Häkkinen, C. Kobel, C. Mateus, Z. Or, J. OReilly, L. Serdén *et al.*, "Diagnosis related groups in europe: moving towards transparency, efficiency, and quality in hospitals?" *BMJ*, vol. 346, 2013.
[2] H. Fahlevi, "The innovation of the role of accounting in public hospitals-lessons learned from hospital financing reforms in indonesia and germany," Ph.D. dissertation, Zugl.: Speyer, Univ., Diss., 2014, 2014.
[3] MENTERI-KESEHATAN, "Peraturan menteri kesehatan republik indonesia nomor 27 tahun 2014."
[4] R. Grishman, "Natural language processing," *Journal of the American Society for Information Science*, vol. 35, no. 5, pp. 291–296, 1984.

[5] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, no. 1, pp. 51–89, 2003.
[6] R. Cuel, A. Francesconi, F. Nardelli, and G. Armellin, "S-grouper a semantic based system to semi-automatic encode hospital activities," in *eKNOW 2015 : The Seventh International Conference on Information, Process, and Knowledge Management*, 2015, pp. 58–59.
[7] M. Miranda and L. Cortez, "The diagnosis related groups (drgs) to adjust payment-mechanisms for health system providers." Conferencia Interamericana de Seguridad Social, 2005.
[8] CDC, "Icd-9 and icd-10 who," http://www.cdc.gov/nchs/icd.htm, accessed: 2016-01-29.
[9] WHO, *International Statistical Classification of Diseases and Related Health Problem Tenth Revision*, 2008.
[10] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms:. the c-value/nc-value method," *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 115–130, 2000.
[11] H. Nakagawa and T. Mori, "A simple but powerful automatic term extraction method," in *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14*. Association for Computational Linguistics, 2002, pp. 1–7.
[12] W. Golik, R. Bossy, Z. Ratkovic, and C. Nédellec, "Improving term extraction with linguistic analysis in the biomedical domain," in *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing13), Special Issue of the journal Research in Computing Science*, 2013, pp. 24–30.
[13] X. Wang and S. Liu, "Analysis and research of enterprise technology competent advantage on text mining and correspondence analysis," *International Journal of Database Theory and Application*, vol. 6, no. 5, pp. 133–140, 2013.
[14] T. Hamon, C. Engström, and S. Silvestrov, "Term ranking adaptation to the domain: Genetic algorithm-based optimisation of the c-value," in *Advances in Natural Language Processing*. Springer, 2014, pp. 71–83.
[15] G. Salton and M. MacGill, "Introduction to modern information retrieval," *McGraw-Hill computer science series*, 1983.
[16] M. van den Branden, N. Wiratunga, D. Burton, and S. Craw, "Integrating case-based reasoning with an electronic patient record system," *Artificial Intelligence in Medicine*, vol. 51, no. 2, pp. 117–123, 2011.
[17] W. Quentin, H. Rätto, M. Peltola, R. Busse, U. Häkkinen *et al.*, "Acute myocardial infarction and diagnosis-related groups: patient classification and hospital reimbursement in 11 european countries," *European heart journal*, vol. 34, no. 26, pp. 1972–1981, 2013.
[18] A. Gojun, U. Heid, B. Weissbach, C. Loth, and I. Mingers, "Adapting and evaluating a generic term extraction tool." in *LREC*, 2012, pp. 651–656.
[19] H.-T. Wang and A. U. Tansel, "Medcase: a template medical case store for case-based reasoning in medical decision support," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 962–967.
[20] M. Chris and H. Schuetze, "Foundations of statistical natural language processing," 1999.

(Check) 24 - A Study and Model of Automatic Mapper For

ORIGINALITY REPORT

%22	%22	%14	%13
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.kela.fi Internet Source	%5
2	Submitted to Colorado Technical University Online Student Paper	%3
3	www.uni-speyer.de Internet Source	%3
4	research.ifmo.ru Internet Source	%1
5	www.iima.org Internet Source	%1
6	www.ncbi.nlm.nih.gov Internet Source	%1
7	Jingjing Kang. "Automatic Domain Terminology Extraction Using Graph Mutual Reinforcement", Lecture Notes in Computer Science, 2010 Publication	%1
8	www.chdl.org Internet Source	%1

9	Benfano Soewito, Ford Lumban Gaol, Echo Simanjuntak, Fergyanto E. Gunawan. "Smart mobile attendance system using voice recognition and fingerprint on smartphone", 2016 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2016 Publication	%1
10	Submitted to Higher Education Commission Pakistan Student Paper	%1
11	www.scribd.com Internet Source	%1
12	www.icd10data.com Internet Source	<%1
13	repub.eur.nl Internet Source	<%1
14	personalpages.manchester.ac.uk Internet Source	<%1
15	www.automation-drive.com Internet Source	<%1
16	ojs.academypublisher.com Internet Source	<%1
17	van den Branden, M.. "Integrating case-based reasoning with an electronic patient record system", Artificial Intelligence In Medicine, 201102	<%1

	Publication	
18	etheses.whiterose.ac.uk Internet Source	<%1
19	www.phila.gov Internet Source	<%1
20	aser.ornl.gov Internet Source	<%1
21	files.opaals.eu Internet Source	<%1
22	www.physiciansadvantage.net Internet Source	<%1
23	www.cdc.gov Internet Source	<%1
24	atrium.lib.uoguelph.ca Internet Source	<%1
25	Wang, Hsien-Tseng, and Abdullah Uz Tansel. "MedCase : a template medical case store for case-based reasoning in medical decision support", Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM 13, 2013. Publication	<%1

EXCLUDE QUOTES ON EXCLUDE MATCHES < 5 WORDS

EXCLUDE ON
BIBLIOGRAPHY